Counterfactual-Guided Diffusion Perception for Multimodal Hate Speech Detection

Nan Mu, Wenzhong Yang*, Yabo Yin, Hongzhen Lv
School of Computer Science and Technology, Xinjiang University, China
{107552304114, 107552304103}@stu.xju.edu.cn, {yangwenzhong, yinyabo}@xju.edu.cn

***Abstract*—Detecting multimodal hate speech is impeded by two coupled challenges: macro-level spurious cross-modal correlations and micro-level noise within fused representations. Existing causal inference methods often overlook the latter, assuming high-quality feature representations despite the inevitable artifacts and background noise introduced during multimodal fusion. To address this, we introduce Counterfactual-Guided Diffusion (CF-DF), a unified framework that synergizes micro-level purification with macro-level debiasing. Specifically, we establish a "Refine-then-Debias" paradigm: at the micro level, a diffusion-based module leverages single-step Tweedie recovery to project noisy fused features back onto the high-density data manifold, thereby enhancing the signal-to-noise ratio. At the macro level, we construct text-absent and image-absent counterfactual paths upon these refined features to disentangle and suppress unimodal shortcuts via causal intervention. Optimized end-to-end, CF-DF demonstrates superior performance and stability compared to state-of-the-art baselines on the MAMI and HarMeme benchmarks, providing a more interpretable and trustworthy solution for multimodal content moderation.**

***Index Terms*—Multimodal hate speech detection, Counterfactual intervention, Diffusion Models, Causal debiasing, Representation Refinement**

I. Introduction

Memes—image–text composites—are now a central medium of online expression and diffusion [1], but also a conduit for hate speech [2]. Detecting multimodal hate is significantly harder than text-only analysis because hateful intent is often conveyed implicitly through the subtle interplay between image and text, heavily relying on culturally situated innuendo [3]. Robust multimodal detectors are therefore essential for safeguarding online platforms.

Prior work has evolved from early feature-concatenation methods [4] to sophisticated architectures leveraging vision–language pre-trained models (VLMs) like CLIP [5]. State-of-the-art models such as HateCLIPper [6] and HyperHatePrompt [7] have achieved competitive performance by designing advanced fusion modules. Concurrently, causal inference has emerged as a powerful tool to mitigate bias. For instance, counterfactual reasoning has been used to address spurious correlations in image recognition and text classification [8], [9], establishing a principled approach for debiasing.

Despite these advancements, two fundamental and coupled challenges remain. The first is confounding bias arising from spurious correlations. Deep models tend to learn "shortcut features"—statistical associations that hold in the training set but fail to generalize (e.g., specific identities frequently co-occurring with hate labels) [10]. The second challenge is the quality of the representation itself [11], [12]. Recent empirical studies indicate that VLM features often suffer from a "modality gap" and noisy correspondence, where the fused representation entangles task-irrelevant background information with semantic content [13], [14]. This is particularly acute in memes, where the image and text are often weakly aligned or ironically juxtaposed.

This creates a critical theoretical misalignment: standard causal interventions operate on the assumption that the underlying representations are robust and disentangled enough to support valid causal estimation [15], [16]. However, given the inherent noise documented in multimodal fusion [13], this assumption is often violated. Consequently, applying causal subtraction directly on contaminated features may lead to imprecise bias estimation, limiting the efficacy of debiasing. Based on this analysis, we propose a "Refine-then-Debias" strategy, positing that representations should theoretically be purified to a high-quality manifold before causal intervention is applied.

To this end, we present Counterfactual-Guided Diffusion (CF-DF). Our approach operates at two complementary levels. **Micro-level Refinement:** We introduce a diffusion module to purify representations in the latent space. Drawing on manifold purification theories [17], we employ an add-noise-then-denoise process with Tweedie's formula adjustment to project noisy fused features back onto the high-density regions of the class-conditional manifold. This step aims to enhance the signal-to-noise ratio, providing a cleaner foundation for causal analysis. **Macro-level Debiasing:** On top of the refined features, we construct a structural causal model with three parallel worlds (factual, text-absent, image-absent). By intervening on these paths, we explicitly calculate and subtract the unimodal shortcut effects from the total effect, seeking to isolate the genuine cross-modal reasoning required for hate detection.

Our main contributions are as follows:

This work is a research achievement supported by the National Natural Science Foundation of China (No.62262065), the National Natural Science Foundation of China (No. 62476233), the "Tianchi Talent" Research Project of Xinjiang (Grant No.51052501821), the "Tianshan Talent" Research Project of Xinjiang (No. 2022TSYCLJ0037)

*Corresponding author.

- We propose CF-DF, a unified framework that synergizes diffusion-based manifold purification with counterfactual causal reasoning, addressing the coupled challenges of representation noise and confounding bias. By explicitly isolating and quantifying the impact of unimodal shortcuts, our approach yields a more interpretable decision-making process, thereby advancing the development of trustworthy and reliable AI systems.
- We design a parallel refinement mechanism using a lightweight diffusion module (approx. 1.2M parameters) that iteratively purifies features along distinct causal paths without imposing heavy computational overhead.
- Extensive experiments on HarMeme and MAMI demonstrate that CF-DF achieves state-of-the-art performance. Crucially, detailed stability analysis (mean $\pm$ std) confirms that our approach exhibits consistent improvements and robustness against random variance.

To increase reproducibility and replicability, the source code and data used in this work will be made publicly available upon acceptance.

II. Methodology

This section elaborates on the proposed method, which is composed of four main components: multimodal feature encoding, macro-level counterfactual intervention, micro-level representation refinement, and causal effect aggregation. The overall framework of our approach is shown in Fig. 1.

A. Multimodal Feature Encoding

For an input (I, T), we first employ the pre-trained vision–language model CLIP [5] to extract image and text features, $f_v \in \mathbb{R}^{d_v}$ and $f_t \in \mathbb{R}^{d_t}$. These features are then aligned to a common d-dimensional space via learnable linear projections:

$$f_i' = W_i f_i + b_i, \quad \forall\, i \in \{v, t\}, \tag{1}$$

where W_i and b_i are learnable weights and biases. These adapters act as task-specific projections that map general-purpose CLIP features into a space tailored to multimodal hate speech detection.

To capture deep cross-modal associations, we design an interaction-based fusion module, $\mathrm{Fusion}_{\mathrm{interact}}(\cdot,\cdot)$, which employs symmetric multi-head cross-attention where each modality queries the other. The final interaction-based representation $F_{\text{interact}} \in \mathbb{R}^d$ is produced as follows:

$$F_{\text{interact}} = W_f[\mathrm{MHA}(f_t', f_v'); \mathrm{MHA}(f_v', f_t')] + b_f, \tag{2}$$

where $\mathrm{MHA}(Q, K, V)$ denotes a standard multi-head attention operation, and W_f, b_f are parameters of the final fusion layer. This interaction-based representation is specifically used to model the total effect in the factual world.

B. Macro-level Debiasing: Counterfactual Intervention

To isolate shortcut effects, we draw on the theory of causal effect decomposition [18]. In multimodal reasoning, the model's prediction (Total Effect, TE) is composed of two parts: the *Natural Direct Effect* (NDE), which represents the contribution from unimodal shortcuts (e.g., specific keywords or visual styles), and the *Natural Indirect Effect* (NIE), which captures the genuine cross-modal interaction we aim to learn. Since NIE cannot be observed directly, we employ counterfactual reasoning to estimate it by subtracting the NDE from the TE. We construct three parallel "worlds" to simulate these effects:

Factual World: The feature $F_o = F_{\text{interact}}$ is generated using the full interaction module defined previously. This path takes the complete input (I, T) and captures the Total Effect, encompassing both genuine reasoning and spurious correlations.

Text-Absent World: To isolate the image's direct effect (visual shortcut), we simulate a counterfactual scenario where the text is blocked. We concatenate the image feature with a zero vector and project back to the target dimension using a shared linear layer:

$$F_{\neg t} = W_{\text{cf}}\,[\,f_v'\,;\,\mathbf{0}_d\,] + b_{\text{cf}}. \tag{3}$$

Image-Absent World: Symmetrically, to capture the text's direct effect (textual shortcut), we simulate the absence of visual information by concatenating the text feature with a zero vector:

$$F_{\neg v} = W_{\text{cf}}\,[\,\mathbf{0}_d\,;\,f_t'\,] + b_{\text{cf}}, \tag{4}$$

where $[\cdot;\cdot]$ denotes concatenation, and $W_{\text{cf}} \in \mathbb{R}^{d\times 2d}$, $b_{\text{cf}} \in \mathbb{R}^d$ are parameters shared by the two counterfactual worlds.

Due to these asymmetric computational paths, the resulting features $\{F_w\}$ may exhibit different statistics. To mitigate this, we apply Layer Normalization independently to each feature, yielding normalized features $\bar{F}_w$ as a stable basis for the subsequent diffusion refinement.

C. Micro-level Purification: Representation Refinement

While the parallel "worlds" separate causal paths, the fused features $F_o, F_{\neg t}, F_{\neg v}$ may still entangle crucial causal signals with bias and task-irrelevant noise [13]. Directly performing causal subtraction on such noisy representations may yield imprecise bias estimation. Therefore, we employ a diffusion model to project potentially off-manifold fused features back onto the high-density regions of the class-conditional manifold (Manifold Projection). This ensures that the subsequent causal intervention operates on semantically valid representations, adhering to the "Refine-then-Debias" paradigm.

1) Diffusion Principles: The forward noising process progressively perturbs clean data x_0 into isotropic Gaussian noise over timesteps $t = 1, \ldots, T$:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{5}$$

where $\bar{\alpha}_t$ follows a predefined noise schedule. The reverse denoising network ϵ_θ is implemented as a lightweight 3-layer

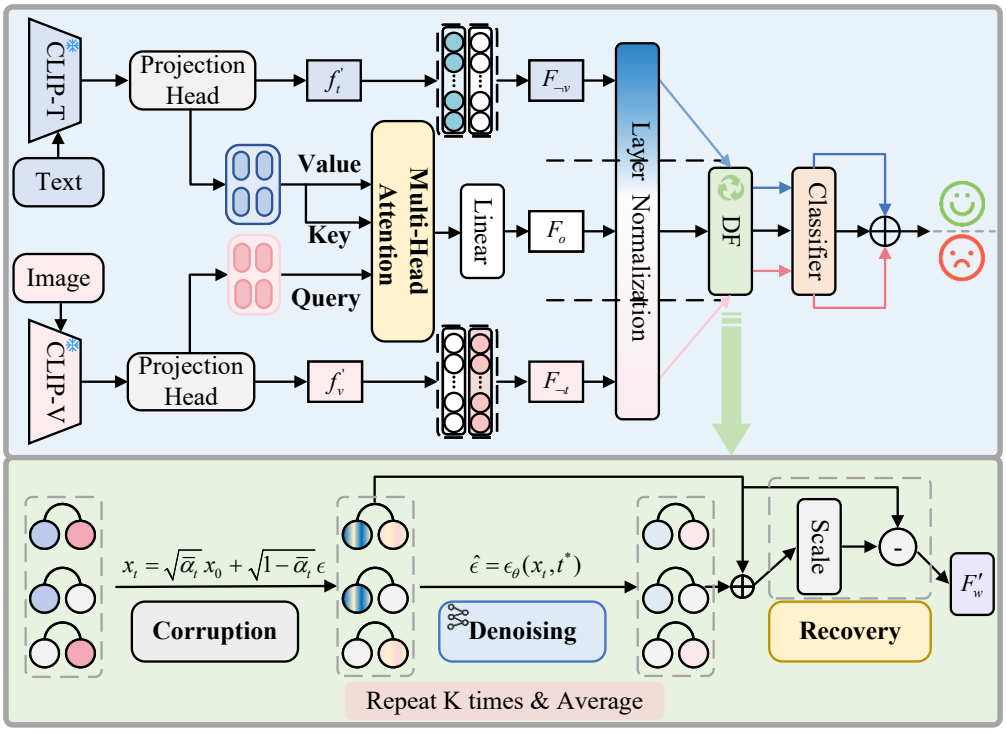

Fig. 1. The overall framework of CF-DF.

MLP with residual connections. It is trained by minimizing the noise prediction error:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,x_0,\epsilon}\Big[\big\|\epsilon_\theta(x_t,t)-\epsilon\big\|_2^2\Big]. \tag{6}$$

2) Application in CF-DF: We treat the fused features from the three "worlds," $\{F_w\}_{w\in\{o,\neg t,\neg v\}}$, as the clean data x_0 to be refined. A single, weight-sharing denoising network ϵ_θ is trained on all branches in parallel, with the total diffusion loss being the sum over the worlds.

At inference, to align the test distribution with the training distribution of the denoising network, we adopt an add-noise-then-denoise refinement strategy. Specifically, we select a small timestep t^* as a hyperparameter and first add mild noise to F_w to obtain $(F_w)_{t^*}$ via (5). We then recover an estimate of the clean feature in a single step using Tweedie's formula [19]:

$$\widehat{F'_w} = \frac{(F_w)_{t^*} - \sqrt{1-\bar{\alpha}_{t^*}}\;\epsilon_\theta\big((F_w)_{t^*},t^*\big)}{\sqrt{\bar{\alpha}_{t^*}}}. \tag{7}$$

This single-step recovery avoids the computational overhead of iterative sampling while effectively projecting features to the manifold. We take $\widehat{F'_o}$, $\widehat{F'_{\neg t}}$, and $\widehat{F'_{\neg v}}$ as the refined representations for the three "worlds."

D. Causal Effect Aggregation and Joint Training

The refined representations $(\widehat{F'_o},\widehat{F'_{\neg t}},\widehat{F'_{\neg v}})$ are fed into a shared classification head to obtain logits for the three branches: $z_o, z_{\neg t}, z_{\neg v}$. Based on the causal decomposition logic ($NIE = TE - NDE$), we seek to isolate the pure interaction signal by subtracting the unimodal NDEs ($z_{\neg t}, z_{\neg v}$) from the total effect (z_o). A stop-gradient operator, $sg(\cdot)$, is applied to the counterfactual terms to ensure they serve purely as bias indicators without affecting the factual representation learning:

$$z = z_o - \alpha\cdot \text{sg}\left(\tanh(z_{\neg t})\right) - \beta\cdot \text{sg}\left(\tanh(z_{\neg v})\right). \tag{8}$$

Here, α and β control the debiasing strength. By explicitly removing the direct unimodal effects, the final logit z approximates the Natural Indirect Effect (NIE), representing the reasoning outcome driven solely by cross-modal interplay rather than unimodal shortcuts.

The final classification loss, $\mathcal{L}_{\text{cls}}$, is the binary cross-entropy (BCE) loss calculated on z. The total training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda\,\mathcal{L}_{\text{diff}}^{(\text{all})}, \tag{9}$$

where λ balances classification and refinement. Notably, this joint optimization endows the diffusion module with an implicit task orientation. While $\mathcal{L}_{\text{diff}}$ regularizes ϵ_θ to learn the intrinsic manifold structure, gradients from $\mathcal{L}_{\text{cls}}$ backpropagate through the refinement pathway to ϵ_θ. This combined signal guides the denoising network to not only preserve manifold consistency but also to adjust features in directions conducive to the final discrimination task, prioritizing the suppression of noise and confounding components detrimental to classification.

III. Experiments

We conduct comprehensive experiments to evaluate the effectiveness, robustness, and efficiency of CF-DF for multimodal hate speech detection.

A. Datasets

We evaluate on two widely used benchmarks: MAMI [20] and HarMeme [21].

MAMI: Derived from SemEval-2022 Task 5, this dataset focuses on misogynistic content, containing 10,000 image–text pairs. We follow the official binary classification setup (misogynous vs. non-misogynous).

HarMeme: This dataset consists of over 8,500 COVID-19-related memes labeled as "very hateful," "partially hateful," or "non-hateful." Following [22], we merge the first two into a single "hateful" category.

B. Baselines

We compare CF-DF with representative text, image, and vision–language baselines, including **BERT-text** [23], **CLIP-text**, **CLIP-image**, **ResNet-image** [24], and CLIP-based multimodal models such as **MOMENTA** [3], **PromptHate** [25], **Pro-Cap** [22], **HyperHatePrompt** [7], **HateCLIPper** [6], and **MemeCLIP** [26]. For fair comparison, we report results from the original papers where the experimental settings align; otherwise, we reproduce the baselines using their official implementations under the same protocol.

C. Implementation Details

We implement CF-DF with frozen CLIP encoders to prevent overfitting; image and text features are linearly projected into a shared 512-dimensional space. We set $(\alpha, \beta) = (0.4, 0.4)$ for MAMI and $(0.3, 0.4)$ for HarMeme, with diffusion loss weight $\lambda = 0.001$. At inference, we average predictions over $K{=}10$ stochastic noise draws. Training uses AdamW [27] (learning rate $1\mathrm{e}{-4}$, weight decay $1\mathrm{e}{-4}$) with linear warm-up, batch size 16, up to 50 epochs, and early stopping with a patience of 10 on validation Macro-F1. All experiments are conducted on a single NVIDIA A40 GPU using PyTorch. The training process takes approximately 3 hours, and peak memory usage is around 4GB, demonstrating the computational efficiency of our lightweight diffusion module. We evaluate performance using Accuracy, Macro-F1, and AUC.

D. Main Results

The results are summarized in Table I. Specifically, we observe the following key findings:

On MAMI, CF-DF attains the highest Accuracy and Macro-F1. Although HyperHatePrompt slightly surpasses CF-DF in AUC (0.843 vs. 0.839), CF-DF leads on correctness-oriented metrics, indicating that counterfactual debiasing with diffusion refinement yields more precise decisions on misogynistic content.

On HarMeme, CF-DF delivers the best Accuracy and Macro-F1, with an AUC comparable to the state of the art. Given the dataset's event- and culture-specific nature, these gains reflect improved robustness and generalization. Counterfactual reasoning reduces reliance on high-frequency pandemic-related cues, while diffusion refinement amplifies genuine image–text interaction signals; their synergy yields notable performance improvements over strong CLIP-based baselines.

TABLE I
Performance comparison of CF-DF with baselines on MAMI and HarMeme. Best results are in bold.

Dataset	Methods	Accuracy	Macro-F1	AUC
MAMI	BERT-text	0.576	0.527	0.661
	CLIP-text	0.619	0.602	0.703
	ResNet-image	0.609	0.587	0.682
	CLIP-image	0.705	0.706	0.812
	MOMENTA	0.721	0.727	0.817
	PromptHate	0.711	0.708	0.808
	Pro-Cap	0.733	0.724	0.832
	HyperHatePrompt	0.753	0.751	**0.843**
	CF-DF	**0.769**	**0.764**	0.839
HarMeme	BERT-text	0.711	0.688	0.763
	CLIP-text	0.738	0.716	0.793
	CLIP-image	0.786	0.784	0.881
	HateCLIPper	0.839	0.833	0.908
	MemeCLIP	0.843	0.834	**0.918**
	CF-DF	**0.854**	**0.846**	0.916

To further verify robustness, we conducted repeated runs with 5 different random seeds. We observed minimal performance fluctuation (standard deviation < 0.01 across all metrics), confirming that the reported gains are consistent and not artifacts of random initialization.

E. Ablation Study

To disentangle the independent contributions of each core component, we evaluate three ablation variants against our full model, with results summarized in Table II.

- **Base Model**: uses only CLIP encoders, projection, fusion, and a classifier.
- **w/o DF**: Retains counterfactual intervention (CF) but removes diffusion refinement (DF).
- **w/o CF**: Discards counterfactual intervention but keeps diffusion refinement on the factual branch.

The results clearly reveal the contribution of each component. Adding either core module to the base model yields substantial gains. Notably, removing counterfactual intervention (w/o CF) leads to a significant drop, consistent with the hypothesis that macro-level intervention effectively limits reliance on unimodal shortcuts. Similarly, removing diffusion refinement (w/o DF) causes comparable degradation, validating that representation purification is crucial for modeling complex cross-modal interactions. The superior performance of the full CF-DF model confirms that macro-level causal debiasing and micro-level representation refinement act synergistically.

TABLE II
ABLATION EXPERIMENTS PERFORMED ON MAMI AND HARMEME DATASETS.

Model configuration	Dataset	Accuracy	Macro-F1
CF-DF	MAMI	**0.769**	**0.764**
	HarMeme	**0.854**	**0.846**
w/o CF	MAMI	0.748	0.742
	HarMeme	0.828	0.821
w/o DF	MAMI	0.753	0.748
	HarMeme	0.835	0.829
Base Model	MAMI	0.729	0.724
	HarMeme	0.815	0.809

F. Parameter Sensitivity and Confidence Analysis

1) Hyperparameter Sensitivity: To assess the robustness of our debiasing coefficients α and β, we conducted a grid search around the optimal settings. As shown in Table III, varying α and β within the range $[0.3, 0.5]$ on HarMeme results in minimal performance fluctuation. This stability indicates that while optimal bias correction strength may vary slightly due to dataset-specific bias patterns, the model is not overly sensitive to precise hyperparameter tuning, easing deployment.

TABLE III
SENSITIVITY OF MACRO-F1 TO α AND β ON HARMEME.

$\alpha \backslash \beta$	**0.3**	**0.4**	**0.5**
0.3	0.844	**0.846**	0.841
0.4	0.842	0.839	0.842
0.5	0.839	0.840	0.841

2) Visualization of Classifier Confidence: To explicitly verify the impact of diffusion-based refinement, we train the same linear probe on frozen features extracted before and after applying DF. Fig. 2 visualizes the class-conditional distributions of $P(\text{hate})$ on the test set.

In Fig. 2(a), the distributions for hate and non-hate classes substantially overlap near the decision boundary ($P \approx 0.5$), indicating that raw fused features often lie in ambiguous regions with high uncertainty. In contrast, Fig. 2(b) shows that after DF refinement, the distributions become sharply bimodal, concentrating efficiently at the extremes ($P \to 0$ and $P \to 1$). This semantic sharpening suggests that the diffusion module successfully pulls noisy features away from the uncertain boundary towards the high-density prototypes of their respective classes (i.e., manifold projection), thereby simplifying the subsequent causal intervention task.

G. Qualitative Analysis

To illustrate how CF-DF addresses the challenges of bias and noise, we present a qualitative comparison with strong baselines in Fig. 3.

In Fig. 3(a), the baseline overfits the textual shortcuts (”black people”, ”doesn’t belong”), misclassifying the meme as Hate despite the benign visual context (a refusal gesture).

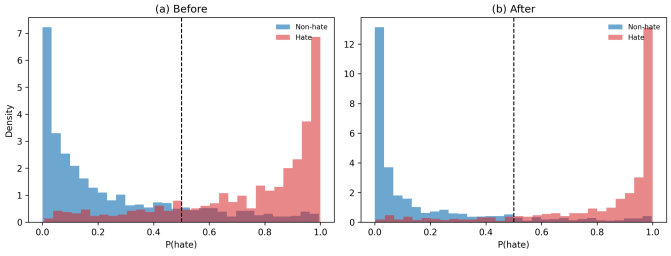

Fig. 2. Probability distributions of $P(\text{hate})$ predicted by a linear probe on features (a) Before and (b) After refinement by our DF module. The dashed line at 0.5 marks the decision boundary.

Fig. 3. Qualitative comparison between baselines and CF-DF. (a) Correction of a False Positive by mitigating unimodal spurious correlations. (b) Correction of a False Negative where diffusion-based refinement enhances implicit semantic signals. Red and Green denote incorrect and correct predictions, respectively.

However, the visual context (refusal) negates the stereotypical interpretation, forming a benign message. CF-DF correctly identifies this as Non-Hate. This correction is attributed to the macro-level counterfactual intervention, which suppresses the high contribution of the textual shortcut path, allowing the model to weigh the visual context more heavily in the final decision.

The right column displays a complex example involving political satire. The text juxtaposes a child’s remark on “spying” with a retort regarding paternity. Recognizing the hateful intent, specifically a conspiracy theory implying pervasive government surveillance, requires bridging the significant semantic gap between these seemingly unrelated concepts. The baseline, potentially hindered by noise in the fused representation, fails to establish this connection, predicting Non-Hate. In contrast, CF-DF successfully detects the hateful intent. This result suggests that the micro-level diffusion module, by projecting features onto the high-density manifold, effectively reduces task-irrelevant noise and enhances the salience of subtle cross-modal cues required to link “surveillance” with the “denial of paternity.”

IV. CONCLUSION

In this paper, we proposed CF-DF, a unified framework designed to address the dual challenges of confounding bias and representation noise in multimodal hate speech detection. By introducing a “Refine-then-Debias” paradigm, we integrate micro-level diffusion-based manifold purification with macro-

level counterfactual intervention. This approach aims to ensure that causal debiasing operates on high-quality features rather than noisy fused representations. Extensive experiments on the MAMI and HarMeme benchmarks demonstrate that CF-DF achieves competitive performance compared to state-of-the-art methods. Crucially, our stability analysis confirms that the model exhibits low variance across random seeds and remains robust under varying hyperparameter settings. Future work will focus on generalizing this causal-diffusion mechanism to more complex multimodal scenarios involving temporal dynamics, further broadening the applicability of trustworthy content moderation.

REFERENCES

[1] S. Zannettou, T. Caulfield, J. Blackburn, E. De Cristofaro, M. Sirivianos, G. Stringhini, and G. Suarez-Tangil, "On the origins of memes by means of fringe web communities," in *Proceedings of the internet measurement conference 2018*, 2018, pp. 188–202.

[2] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," *Advances in neural information processing systems*, vol. 33, pp. 2611–2624, 2020.

[3] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, and T. Chakraborty, "Momenta: A multimodal framework for detecting harmful memes and their targets," *arXiv preprint arXiv:2109.05184*, 2021.

[4] S. Sai, N. D. Srivastava, and Y. Sharma, "Explorative application of fusion techniques for multimodal hate speech detection," *SN Computer Science*, vol. 3, no. 2, p. 122, 2022.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[6] G. K. Kumar and K. Nandakumar, "Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features," in *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, L. Biester, D. Demszky, Z. Jin, M. Sachan, J. Tetreault, S. Wilson, L. Xiao, and J. Zhao, Eds. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 171–183. [Online]. Available: https://aclanthology.org/2022.nlp4pi-1.20/

[7] B. Xu, E. Yu, J. Zhou, H. Lin, and L. Zong, "Hyperhateprompt: A hypergraph-based prompting fusion model for multimodal hate detection," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 3825–3835.

[8] J. Sun, M. Gao, H. Wang, and Q. Dong, "Recursive counterfactual deconfounding for image recognition," *Knowledge-Based Systems*, vol. 315, p. 113245, 2025.

[9] Z. Chen, L. Hu, W. Li, Y. Shao, and L. Nie, "Causal intervention and counterfactual reasoning for multi-modal fake news detection," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 627–638.

[10] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[11] H. Lv, W. Yang, Y. Yin, F. Wei, J. Peng, and H. Geng, "Mdf-fnd: A dynamic fusion model for multimodal fake news detection," *Knowledge-Based Systems*, vol. 317, p. 113417, 2025.

[12] L. Zhang, X. Zhang, Z. Zhou, F. Huang, and C. Li, "Reinforced adaptive knowledge learning for multimodal fake news detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 15, 2024, pp. 16 777–16 785.

[13] W. Liang, Y. Zhang, Y. Kwon, S. Ye, and J. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 17 612–17 625.

[14] Z. Huang, G. Zeng, B. Liu, J. Fu, and J. Fu, "Learning with noisy correspondence for cross-modal matching," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 622–633.

[15] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

[16] R. Suter, D. Miladinovic, B. Schölkopf, and S. Bauer, "Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6056–6065.

[17] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, "Diffpure: Diffusion models for adversarial purification," in *International Conference on Machine Learning*. PMLR, 2022, pp. 16 069–16 084.

[18] J. Pearl, "Causal inference in statistics: An overview," 2009.

[19] B. Efron, "Tweedie's formula and selection bias," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1602–1614, 2011.

[20] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, and J. Sorensen, "Semeval-2022 task 5: Multimedia automatic misogyny identification," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 2022, pp. 533–549.

[21] S. Pramanick, D. Dimitrov, R. Mukherjee, S. Sharma, M. S. Akhtar, P. Nakov, and T. Chakraborty, "Detecting harmful memes and their targets," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 2783–2796. [Online]. Available: https://aclanthology.org/2021.findings-acl.246/

[22] R. Cao, M. S. Hee, A. Kuek, W.-H. Chong, R. K.-W. Lee, and J. Jiang, "Pro-cap: Leveraging a frozen vision-language model for hateful meme detection," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 5244–5252.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] R. Cao, R. K.-W. Lee, W.-H. Chong, and J. Jiang, "Prompting for multimodal hateful meme classification," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 321–332. [Online]. Available: https://aclanthology.org/2022.emnlp-main.22/

[26] S. B. Shah, S. Shiwakoti, M. Chaudhary, and H. Wang, "MemeCLIP: Leveraging CLIP representations for multimodal meme classification," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 17 320–17 332. [Online]. Available: https://aclanthology.org/2024.emnlp-main.959/

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*.